

On the Importance of Tactile Sensing for Imitation Learning: A Case Study on Robotic Match Lighting

Niklas Funk¹, Changqi Chen¹, Tim Schneider¹, Georgia Chalvatzaki¹, Roberto Calandra², Jan Peters¹

Abstract—The field of robotic manipulation has advanced significantly in the last years. At the sensing level, several novel tactile sensors have been developed, capable of providing accurate contact information. On a methodological level, learning from demonstrations has proven an efficient paradigm to obtain performant robotic manipulation policies. The combination of both holds the promise to extract crucial contact-related information from the demonstration data and actively exploit it during policy rollouts. However, despite its potential, it remains an underexplored direction. This work therefore proposes a multimodal, visuotactile imitation learning framework capable of efficiently learning fast and dexterous manipulation policies. We evaluate our framework on the dynamic, contact-rich task of robotic match lighting - a task in which tactile feedback influences human manipulation performance. The experimental results show that adding tactile information into the policies significantly improves performance by over 40%, thereby underlining the importance of tactile sensing for contact-rich manipulation tasks. Project website: <https://sites.google.com/view/tactile-il>.

I. INTRODUCTION

Robotic manipulation remains far from matching the dexterity and efficiency of human hands [1], [2]. In fact, the current trend of exploiting human demonstration data for learning robotic manipulation [3], [4], [5] actively exploits human task understanding and their advanced manipulation capabilities. While it is well-known that human manipulation heavily benefits from touch sensing [6], the majority of current works in imitation learning for manipulation are still missing out on this modality [4], [5], [7]. Given the importance of touch for human manipulation, the question arises *whether robotic policies could also benefit from adding tactile sensing*.

This work approaches this question by studying the impact of touch sensing for learning a dynamic task, namely, igniting matches. We argue that match lighting is an effective testbed for examining the role of touch sensing in learning robotic manipulation from demonstrations. This is because the task requires dynamic motion and compliance [8], which introduces additional complexity compared to standard tasks such as pick-and-place or insertion [9], [10]. Moreover, it is a task for which there is evidence that the availability of touch sensing impacts human performance [11]. Despite the task’s relevance, to the best of our knowledge, it has only been investigated previously in [8]. Yet, Kronander et al. [8] considered fixed match grasp poses and a precisely

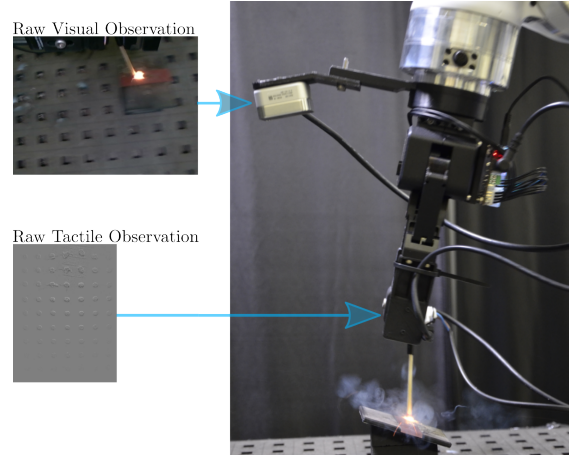


Fig. 1: Autonomous rollout of a policy that is conditioned on visual and tactile observations illustrated on the left. The policy controls the robot and, thereby, the contact configuration between the match and striker paper. As can be seen, the policy ensures sufficient force and velocity, resulting in successfully igniting the match. This work highlights the importance of tactile sensing for reliably solving the dynamic and delicate task of lighting up matches.

calibrated setup without including high-dimensional observations. Our work addresses more complicated scenarios, including varying grasp poses and striker paper orientations, while considering RGB camera images, the end effector velocity, and the information from an event-based optical tactile sensor as observations (cf. Fig. 1).

We propose a multi-modal learning from demonstrations framework to solve this intricate manipulation task solely from local embodied sensing. To further restrict the human efforts for learning the task, we emphasise learning from a few demonstrations and consider only 20 available demonstrations. This data is then exploited to learn an expressive multi-modal flow matching policy [12] suitable for reactivity and real-time inference. Given this low data regime, we employ a modular and efficient policy architecture that allows us to compare different encoding and training strategies given the real-world observation data. The experiments demonstrate the efficiency of the proposed framework and showcase that the visuotactile policies can robustly light up matches across different scenarios and observation-encoding strategies despite learning from only 20 demonstrations. They also reveal that the vision-only policies perform considerably worse throughout all evaluations. Additionally, we find that vision-only policies can benefit from employing a

¹Technical University of Darmstadt, Darmstadt, Germany

²LASR Lab, TU Dresden, Dresden, Germany

Corresponding author: Niklas Funk. Email: niklas@robot-learning.de

masked training procedure that exploits tactile observations during training. The results, therefore, underline that tactile information is a crucial source of information for obtaining reliable robotic match lighting policies.

Overall, we contribute a multi-modal framework for efficiently learning robust and reliable manipulation policies suitable for dynamic tasks such as lighting up matches. Moreover, we present a masked training procedure that exploits the tactile signals during training and allows for increased success rates of vision-only policies. Lastly, we contribute an extensive evaluation conducted in our modular real-world match-lighting testing environment. The experiments across different policies and experiment configurations highlight that tactile observations are crucial for obtaining performant policies for dynamic tasks like match lighting and closely matching the human demonstration data.

II. RELATED WORK

Artificial tactile sensors are a promising technology to advance robotic manipulation as they allow direct sensing of the contact configuration between the robot and its environment [13]. Together with the emergence of commercially available [14] and open-source tactile sensors [15], [16], the field of tactile robotic manipulation is gaining increased attention.

One popular approach to obtain tactile manipulation policies is through reinforcement learning [17], [18], [19]. Since reinforcement learning requires exploration, learning performant policies demands a vast amount of environment interactions. To account for this issue, previous works rely on simulation, allowing for fast sample generation while at the same time mitigating the sim-to-real gap [18], [19], [20], [21]. Alternatively, [17], [22] presented approaches for learning policies directly on real robots. However, this requires a carefully designed experimental setup allowing for autonomous exploration, as multiple hours of real-world interactions are needed for successful policy learning. Since our task of match lighting is challenging to simulate, and since safety considerations hinder realizing autonomous exploration on the real system, this work takes a different direction. We want to efficiently learn match lighting policies from few real-world expert demonstrations, thereby significantly reducing the data requirements.

The field of learning robotic manipulation policies from demonstration data [3], [23] has lately received increasing attention [4], [5], [7], [24], [25]. In particular, several works showed the effectiveness of training generative models based on expert demonstrations for obtaining advanced real-world manipulation skills [4], [5]. The field also benefits from efforts for building effective devices for collecting human demonstrations [26], [27]. However, the majority of works in imitation learning focus on quasi-static manipulation tasks and only incorporate RGB or RGBD cameras as external sensors without considering tactile information [4], [5], [7], [26]. This work follows the current efforts and proposes an efficient and modular multi-modal framework for learning from demonstrations by leveraging a generative model trained as a policy. Yet, it differs in that it considers tactile sensors as

input modality and investigates the contact-rich and dynamic manipulation task of igniting matches. Only more recently, a few works investigated adding tactile sensing capabilities [9], [10], [28] into imitation learning frameworks. Yet, these works also focused on quasi-static manipulation tasks and did not consider dynamic manipulation as we do herein. Moreover, this work additionally introduces a masked training procedure, showcasing that considering tactile observations during training can enhance the inference performance of vision-only policies.

From a task-level perspective, [8] is closest related as it also investigates learning match lighting policies from human demonstrations. To achieve good task success rates, they propose employing a varying stiffness controller learned through information from a human-robot interface. Instead of learning a variable stiffness controller, this work directly learns a reactive policy capable of controlling the contact forces by varying the desired target poses. Moreover, this work extends upon [8] in that it considers a more realistic experimental setup, including varying match poses, striker paper orientations, and conditioning the policies onto high-dimensional image and tactile observations.

Overall, we contribute a framework for learning visuo-tactile robotic match lighting policies from human demonstrations and showcasing that tactile sensing is crucial for learning high performance policies on this dynamic task.

III. LEARNING MATCH LIGHTING POLICIES FROM DEMONSTRATIONS

This section describes our approach for learning the dynamic manipulation skill of lighting up matches from few real-world expert demonstrations and deploying the policies on the real system. In terms of sensing, this work exclusively considers local, embodied information, i.e., the image information from an Intel RealSense D405 camera mounted in the robot’s wrist, an open source Evetac [16] tactile sensor mounted within the parallel gripper, and local velocity information (cf. Fig. 2). The following sections detail the learning framework, the policy architecture, the data collection, and the policy inference procedure.

A. Fast and Reactive Multi-modal Policies through Conditional Flow Matching

Our multimodal policy learning framework leverages a generative model as policy. Given the current observations, the generative model should output an action sequence that is close to the demonstrations. Since the match lighting task is delicate and requires reactivity, we propose to learn a policy using flow matching [29]. In particular, we learn an SE(3)-Rectified Linear flow model [7] that generates high-quality samples within low inference times. We impose a flow in SE(3), as the model’s output should be the desired future trajectory of the robot end-effector. In other words, the resulting policies’ action space is a sequence of $N = 16$ SE(3) poses, $T_a = (T_a^1, \dots, T_a^N) \in SE(3)^N$.

By design, the datapoint-conditioned SE(3)-Rectified Linear flow $\phi_t(a|a_1)$ connects an initial noisy sample

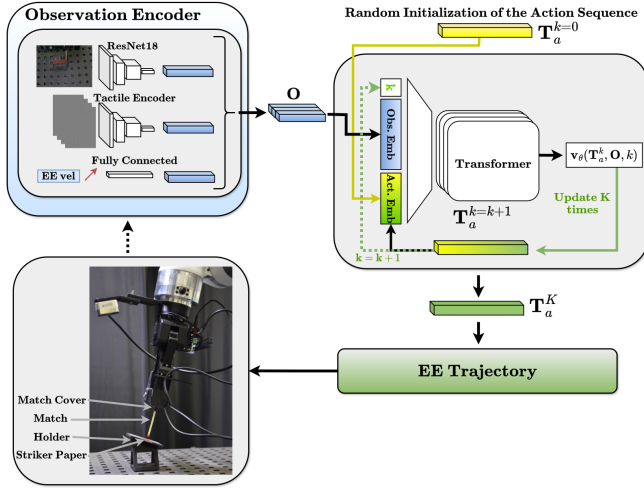


Fig. 2: Method Overview. Upon retrieving the current observations, they are first encoded individually inside the observation encoder and brought into a common shape, i.e., each modality contributes a latent vector of a fixed shape. These latent vectors, together with the current action sequence & time index, then serve as the input to the transformer architecture, which outputs velocities to iteratively refine the action sequence through flow matching. Upon retrieving the final desired end effector trajectory, it is sent to the robot and tracked through a Cartesian Impedance Controller. Note that we only apply the first two actions of the sequence to maintain reactivity.

$\mathbf{a}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ at $t=0$ to a sample from the dataset $\mathbf{a}_1 \in \mathcal{D}$ at $t=1$ with a straight line path. For explanation purposes, in the following, we will consider only a single SE(3) action pose ($\mathbf{p}_1 \in \mathbb{R}^3, \mathbf{r}_1 \in SO(3)$) that is to be generated starting from an initially randomly sampled pose ($\mathbf{p}_0 \in \mathbb{R}^3, \mathbf{r}_0 \in SO(3)$). When decoupling the translational (\mathbf{p}) and rotational (\mathbf{r}) component of the flow, we obtain $\dot{\mathbf{p}}_t = (\mathbf{p}_1 - \mathbf{p}_t)/(1-t)$ & $\dot{\mathbf{r}}_t = (\text{Log}(\mathbf{r}_t^{-1}\mathbf{r}_1))/(1-t)$ for the flow’s velocity field. Therefore, this vector field describes how the current noisy pose needs to be refined to reach the sample from the dataset. Given the training dataset from the demonstrations, the objective is then to train a parameterized Flow Matching model $\mathbf{v}_\theta(\mathbf{p}_t, \mathbf{r}_t, \mathbf{O}, t)$, that, conditioned on the current observation \mathbf{O} and “action” pose, outputs translation and rotation velocities ($\mathbf{v}_p \in \mathbb{R}^3$ & $\mathbf{v}_r \in \mathbb{R}^3$) matching the Flow Matching targets. The model is trained by minimizing $\mathcal{L} = \|\mathbf{v}_p - \dot{\mathbf{p}}_t\|^2 + \|\mathbf{v}_r - \dot{\mathbf{r}}_t\|^2$. Given the learned model, during inference, we sample actions by iteratively refining random initial actions through $\mathbf{p}_{k+1} = \mathbf{p}_k + \mathbf{v}_\theta(\mathbf{p}_t, \mathbf{r}_t, \mathbf{O}, t)\Delta t$ & $\mathbf{r}_{k+1} = \mathbf{r}_k \text{Exp}(\Delta t \mathbf{v}_\theta(\mathbf{p}_t, \mathbf{r}_t, \mathbf{O}, t))$.

B. Policy Architecture

As described in the previous section, our approach employs a parameterized SE(3)-Rectified Linear Flow matching model for obtaining action trajectories. At the core of this policy is a multimodal transformer architecture that

receives observations from multiple sensors, including the RGB camera image, the current end-effector velocity, and, when available, observations from the Evetac tactile sensor as input. Transformers are particularly suitable for this task as they can seamlessly handle the multiple multimodal input observations [30]. The resulting transformer-based policy architecture is illustrated in Fig. 2.

The observations are the crucial source of information for refining the actions. Since we later want to compare different sensor combinations, we ensure modularity, i.e., the individual observation modalities are first encoded individually into latent vectors of dimension 64. We want to point out that the first 5 entries of this 64-dimensional vector are learnable weights that should inform the transformer about the type of observation modality. These latent vectors then serve as the input to a transformer for refining the action sequence. Importantly, the latent observations and entries of the action sequence enter the transformer as individual tokens. The modular policy architecture thus allows for seamlessly evaluating the policies’ performance under different observation encoders. It also enables a masked training procedure that stochastically decides upon the modalities which are available in the transformer. The image observations are processed through a pre-trained ResNet 18 [31] or by training the ResNet from scratch. For the tactile observations, we consider the pre-trained model from [16], and training this architecture from scratch. These features (i.e., one per observation modality, one for each action in the action sequence, and one for the current time index) are the inputs to the transformer model, which consists of 4 layers with 4 attention heads. Inside the transformer, the inputs exchange information with each other and update their embeddings through multi-head attention [32]. In its standard implementation, all the inputs exchange information with each other (including self-connections). Herein, we configure the transformer’s attention mask to full connectivity between the observation tokens, while the action tokens solely cross-attend to the observation tokens. The value of the action tokens thus does not influence the update of the observation tokens. This choice is made because only the observations contain information on how to update the action sequence, while the action sequence only contains noise, especially at the beginning. Moreover, the self-attention within the action tokens is configured such that action poses in the sequence only attend to previous actions. In addition to this masking scheme regarding the actions, in the experiments, we will also investigate the effectiveness of employing stochastic masking at the observation level during training. In particular, we will train a single transformer model that is provided with tactile observation during training with a probability of 50%. Due to this stochasticity on the input level, the policy, therefore, has to better align the latents of the vision and touch observations so that it can generate good outputs in both cases, i.e., when touch is available and when it is not.

The transformer’s final output is the updated action features representing the velocity vectors for the iterative refinement, which is repeated $K=5$ times. Note that the

observations only need to be encoded once. After obtaining the final action sequence, it is sent to the controller and applied to the robot. Using this generative model as policy yields online action generation as illustrated in Fig. 2.

C. Data Collection

Similar to [8], we collect the demonstrations through kinesthetic teaching. This procedure ensures that the human demonstrator directly feels the interaction forces between match and striker paper. This has been crucial for realizing high task success rates during data collection. From a task-level perspective, to light up the match, the match tip must first be brought into contact with the striker paper. Subsequently, the match tip has to be moved along the striker paper while applying sufficient force with sufficient velocity.

Figs. 1 & 2 depict the components of our real-world match lighting environment. Throughout the entire demonstration, we record all of the sensor data, i.e., the image from the wrist-mounted Intel RealSense D405 camera, an open-source Evetac [16] tactile sensor mounted within a Robotis RH-P12-RN gripper attached to the end effector of a 7-DoF Franka Panda, and the local end-effector velocity information. Moreover, we also record the end-effector poses that the robot moves through, as these contain the trajectory information that the robot should follow. Yet, we want to emphasize that the policy framework only operates on the level of local poses expressed in the current end effector frame. While Evetac naturally returns asynchronous event information, for compatibility with the other sensors, we convert the event information into image form. This is done by accumulating the events for every pixel for a duration of 40 ms. In line with this choice, we also collect all the other sensor information at 25 Hz. Since the task is delicate, image (or tactile image) resolution might be crucial. Thus, we maintain a high resolution of 320×240 pixels. As shown in Figs. 1 & 2, for the image observations of the wrist-mounted camera, we ensure that the match and, in particular, the tip of the match is fully observable during the trajectories. Moreover, we found that using the striking surfaces of regular paper matchboxes resulted in short durability after a few experiments. We, therefore, decided to 3D-print a thin rectangular plate to hold the striker paper. In its standard configuration, the plate is raised and placed with an angle of 20° relative to the table (cf. Fig. 1). We used long standard matches with dimensions of $(100\text{ mm} \pm 5\text{ mm}) \times (4\text{ mm} \pm 1\text{ mm}) \times (4\text{ mm} \pm 1\text{ mm})$ to keep the fire at a sufficient distance from the silicone surfaces of the tactile sensors mounted inside the gripper. Lastly, we also 3D printed hollow cylindrical cones to cover the upper 45 mm of the matches. This was necessary to significantly increase the longevity of the silicone gels that cover the tactile sensor, which could rip easily when in direct contact with the matches.

D. Policy Inference and Robot Control

We use the Cartesian Impedance Controller from [33] to move the robot during the autonomous policy rollouts. We tuned the controller’s stiffness and damping values on a few

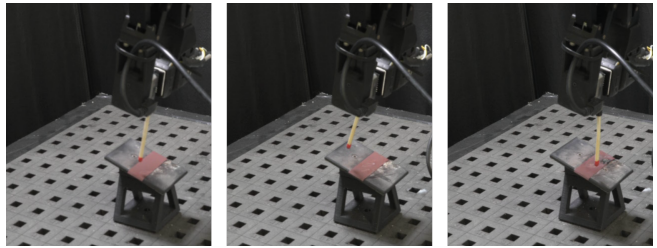


Fig. 3: Visualizing the versatility of the initial configurations during the experiments. Left: Fixed grasp pose strategy. Middle & Right: Two examples of the variable grasp initialization. Note how the initializations yield different configurations w.r.t. distance and angle between match and striker paper that the policies have to handle for solving the task.

of the collected demonstration trajectories. The gains have been chosen such that replaying the trajectories obtained during kinesthetic teaching yields task success when tracked using this control strategy. We rely on the Robotic Operating System (ROS) to gather the sensor observations. Policy inference is run asynchronously, and only the first two actions of the action sequence are applied by the controller before updating the action sequence based on the most recent model inference with the latest observations. The resulting policies run online in real-time as action generation, i.e., policy inference, only takes 0.018 s for our largest vision+touch policies on an NVIDIA RTX 3090 GPU.

IV. EXPERIMENTAL RESULTS

This section evaluates our proposed approach. It is structured along the following three main questions to investigate the importance of tactile sensing for the dynamic manipulation task of lighting up matches: **A:** How important is tactile feedback for obtaining performant match lighting policies? **B:** Can the vision-only policies benefit from leveraging the tactile information during training?, and **C:** Are the policies robust w.r.t. generalizing to novel scenarios?

The following evaluation considers two task versions. One in which the match is always grasped with the same pose, and a more complicated one, where the grasping location is varied within translational offsets of $\pm 1\text{ cm}$ & rotational perturbations of $\pm 10^\circ$ (cf. Fig. 3). For both tasks, we collected 20 successful demonstrations within 1 hour. We then trained our models for 500 epochs. The evaluations report the mean performance across task and model configurations. We trained 3 seeds per combination and evaluated the last checkpoint through 5 rollouts on the real system.

A. How important is tactile feedback for obtaining performant match lighting policies?

Fixed Grasp Pose. In the fixed grasp pose scenario (cf. Fig. 3, left), the vision+touch policies outperform the vision-only policies, achieving a success rate of 86% compared to 33%. Apart from the differences in success rate, Fig. 4 also reveals that the rollouts of the vision+touch (also referred to as visuotactile) policies better match the demonstration data.

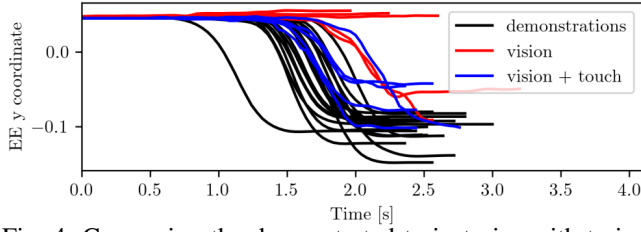


Fig. 4: Comparing the demonstrated trajectories with trajectories obtained from rolling out different policies, considering the y-coordinate of the end effector. The y-coordinate is the direction in which the robot needs to accelerate to light up the matches along the striker paper. Qualitatively, the vision+touch policies generate rollouts that better match the demonstration data compared to the behaviour of the vision-only policies, indicating that the tactile observations contain crucial information for explaining and matching the human demonstrations.

In particular, the visuotactile policy evaluations better align in terms of the timing of accelerating along the striker paper, which corresponds to the end-effectors y-axis. This finding hints that vision-only policies struggle to precisely detect the point in time of making contact since this indicates that the acceleration phase along the striker paper should follow.

Variable Grasp Pose. We repeat the procedure for the variable grasping poses (cf. Fig. 3, middle & right), yet considering a wider class of observation encoders. In particular, we train policies with the pre-trained encoders and either freeze or optimize them during policy training. We also investigate training the observation encoders from scratch. As presented in Fig. 5, in this new, more complicated scenario, there remains a significant difference between the vision-only and vision+touch policies in terms of success rate. Importantly, the superior performance of the visuotactile policies holds across the observation encoding strategies, and adding the tactile observations improves the task success rates by at least 50%. While the best visuotactile policies achieve an average success rate of 80%, the best performing vision-only policies only reach success rates of up to 20%.

Fig. 6 provides a more detailed comparison in that it differentiates between different failure modes of the policies. We consider four types of failures: 1) making contact in the wrong location, i.e., the tip of the match not making contact with the striker paper, 2) not making contact at all during the policy rollout, i.e., the policy accelerating along the striker paper but without making contact, 3) insufficient contact force, i.e., making contact in the right location but without sufficient force resulting in the match not lighting up, and 4) applying too much force during the rollout, i.e., the policy pressing the tip of the match with too much force against the striker paper which results in the match sliding through the fingers. The last failure case is mainly related to the policy missing the transition between the approaching phase of the task and the phase of accelerating along the striker paper. As shown in the comparison, the vision-only policies exhibit significantly increased failure rates. In particular, the failures

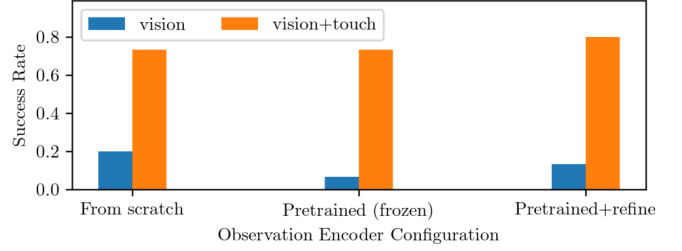


Fig. 5: Comparing the success rates of different policies on the variable grasp pose task. Across different observation encoding strategies, the vision+touch policies consistently outperform the vision-only policies by at least 50%, thereby highlighting the importance of tactile sensing for obtaining reliable match lighting policies.

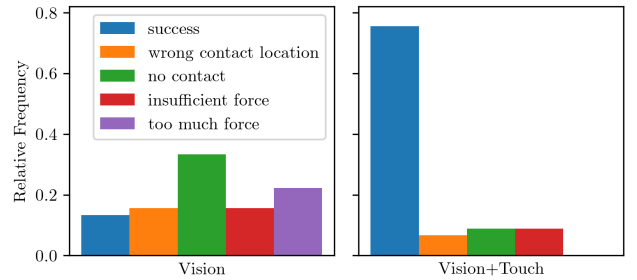


Fig. 6: Comparing success rates and different failure modes for the vision-only and vision+touch policies in the variable grasp pose evaluation. The results are averaged across the observation encoder configurations (cf. Fig. 5). The vision+touch policies have a reduced failure rate of more than 50% and reduce the contact-related failures of not applying any force, insufficient force, or too much force significantly.

related to resolving the current contact state (no contact, insufficient force & too much force) are the most prominent ones with 33%, 15%, & 22%, respectively. In contrast, adding the tactile observations yields significantly reduced failure rates. The few failure cases of the vision+touch policies are mainly related to not making contact at all or not applying sufficient contact forces (8% each), while none of the vision+touch policies apply too much contact force.

Lastly, Fig. 7 illustrates the evolution of the attention weights of the individual inputs of the transformer w.r.t. the update of the fifth action of the action sequence for a visuotactile policy (trained using pre-trained weights but further refining the encoder during training). In other words, it shows how the inputs contribute to updating the action. As can be seen, initially, the vision observation from the RealSense is the most important modality. This is expected, as the camera information is crucial to moving the robot closer to the striker paper. The event-based tactile sensor does not provide any information during this phase, as there is no change in contact configuration. However, once contact is made, the tactile inputs gain importance and become the

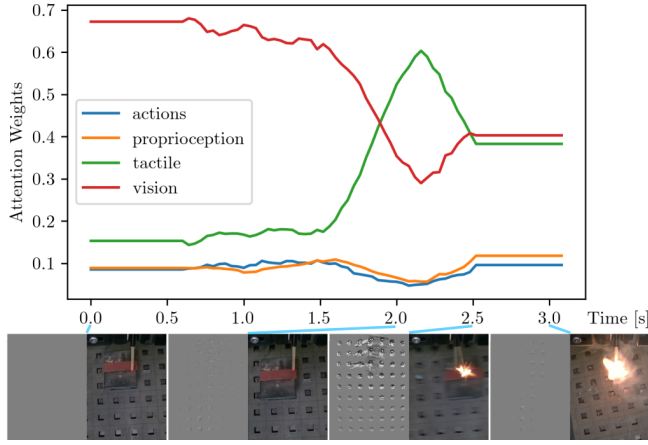


Fig. 7: Visualizing the evolution of the attention weights over time for one exemplary trajectory. The bottom images show the task progression. The plot shows the weights that are attributed to the individual inputs of the transformer: 1) the actions, 2) the proprioception observation (end effector velocity), 3) the tactile observation from Evetac, and 4) the vision observation from the Realsense camera. The weights are w.r.t. to updating the fifth action of the desired end-effector trajectory, which is computed for every observation along the rollout. At the beginning and end of the trajectory (when there are no tactile signals), vision is the most important modality. Once there are changes in contact configuration, touch is the most important modality for action generation, therefore highlighting that touch provides important feedback for controlling the contact configuration.

most important entity. This holds true until the match ignites, which signals successful task execution. The other inputs, i.e., attention to the other actions and to the proprioception observation, stay low throughout the trajectory.

Overall, based on the findings from these experiments, we conclude that touch is a crucial sensing modality for learning performant match lighting policies from few demonstrations.

B. Can the vision-only policies benefit from leveraging tactile information during training?

While the previous section showed the importance of conditioning the policies onto tactile signals, this section investigates whether vision-only policies can benefit from leveraging tactile information during policy training. In particular, we exploit the transformer architecture’s natural capability to handle input sequences of different lengths and investigate the effectiveness of the masked training procedure (cf. Sec. III-B). During the masked training, the policy either receives all of the input modalities or all of the input modalities except for the tactile signals. The masking probability is set to 50%. Since the policy uses the same transformer independent of the masking, it has to align the latent spaces to generate meaningful outputs given the different input combinations. This experiment now investigates whether the masked training procedure can improve the performance of the vision-only policies in the

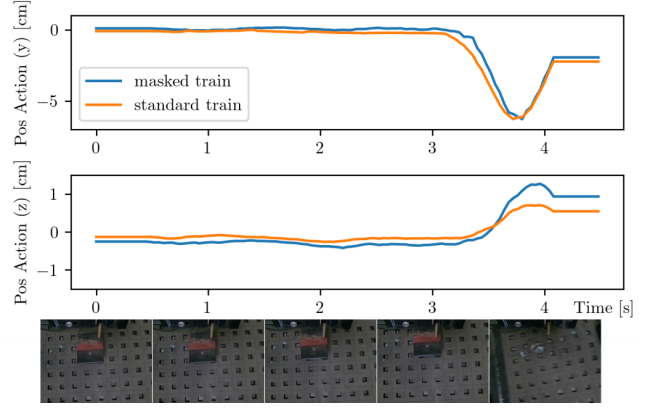


Fig. 8: Comparing the policy predictions of two vision-only policies (one trained with the standard procedure, the other one with the masked one). We visualize the policy predictions for the y- and z-component of the 5th action in the sequence on a trajectory that was obtained by rolling out the standard policy and on which it fails to establish sufficient contact between the match and the striker paper. As shown, the policy that underwent the masked training procedure proposes different actions, i.e., moving closer to the striker paper before accelerating along the striker paper (as shown for the z-predictions when $T < 3.5$ s). Additionally, it proposes to accelerate at a later point in time along the striker paper, as shown for the y-axis predictions.

TABLE I: Success Rate of different vision-only policies in the variable grasp scenario. The policies differ regarding the training procedure, i.e., whether they are trained with the standard procedure or with masked training that considers the tactile signals during training. The masked training procedure, i.e., leveraging touch during training, is effective and yields increased success rates.

	Training Configuration	
	Standard Training	Masked Training
Success Rate	20%	40%

variable grasp pose scenario. We start with the pre-trained encoders and optimize them during the training. This choice is made because the pre-trained encoders already provide a meaningful embedding when the respective modality is important. This is particularly important for the tactile representation, as the masked training procedure indirectly forces the optimization of the vision encoder to account for the missing tactile information.

As shown in Tab. I, the vision-only policies that have undergone the masked training procedure achieve significantly higher success rates, increasing the number of successful rollouts by a factor of 2 and achieving an overall success rate of 40%. In particular, while the policies trained with the standard procedure often fail to establish contact between the match and the striker paper (46% in this experiment), the policies that underwent the masked training procedure exhibit a significantly decreased probability of this failure

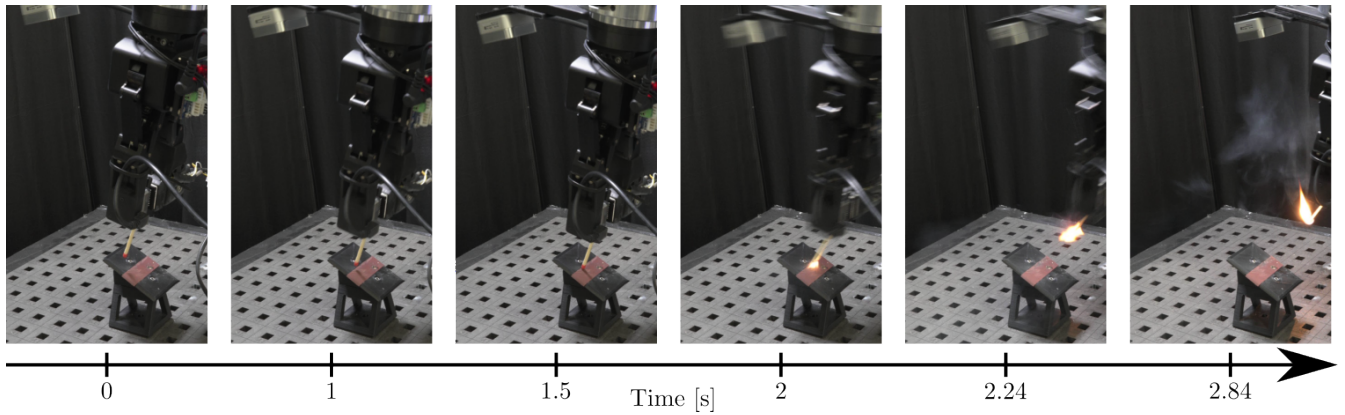


Fig. 9: Successful policy rollout for the out-of-distribution evaluation. Even though the demonstrations are given when the angle between the striker paper and the table is 20° , our policy shows generalization to a mounting angle of 30° .

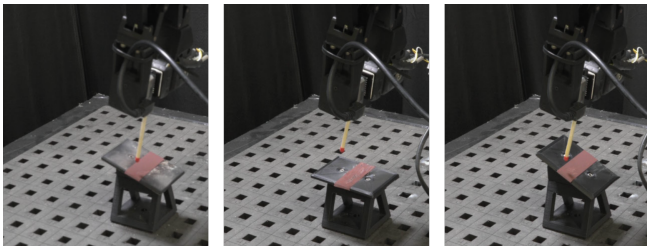


Fig. 10: Visualizing the different mounting angles of the striker paper in the generalization experiments. Left: Nominal mounting angle of 20° . Middle & Right: Mounting angles used in the generalization experiments of 5° and 30° , respectively. Note how the different mounting angles change the angle & distance between match and striker paper.

mode (25%). To underline this finding quantitatively, Fig. 8 compares the differently trained policies regarding action generation. It visualizes the translational outputs for the 5th action in the sequence along the y- (direction of acceleration along the striker paper) and the z-direction (controlling the height of the match tip). For the comparison, we consider a trajectory that has been obtained by rolling out the policy trained with the standard procedure. During this trajectory, the policy failed to establish contact between the match and the striker paper. Considering the z-component of the predicted action, before the start of the sideways motion, the policy that was trained using the masked procedure outputs lower values, thereby indicating that it wants to move the end effector lower, increasing the probability of making contact with the striker paper. Considering the y-direction, it is also evident that the policy trained using the standard procedure aims to move along the striker paper earlier. This behaviour again increases the probability of accelerating too early without making proper contact with the striker paper. We conclude that the masked training procedure increases the success rates of vision-only policies. Therefore, the availability of tactile observations can improve policy performance, even when tactile feedback is only provided during training.

C. Are the policies robust w.r.t. generalizing to novel scenarios?

This last experiment evaluates whether the visuotactile policies can generalize to novel scenarios in which the angle between the match and the striker paper is further changed compared to the previous experiments and demonstrations. This is achieved by mounting the striker paper at novel, previously unseen mounting angles of 5° and 30° (cf. Fig. 10), while maintaining the variable grasping. This evaluation considers the visuotactile policy with the pretrained+refine training procedure.

Fig. 9 depicts a successful policy rollout in one of the novel environments. Quantitatively, for the two testing scenarios, we obtain success rates of 80% for the 5° mounting angle and 67% for 30° . The policies, therefore, generalize without losing performance to the 5° mounting angle. One explanation for the slightly decreased success rates for the 30° mounting angle is that the match starts closer to the 3D-printed holder, leaving the policies less space to adjust the angle between the match and the striker paper. Providing more demonstrations for this specific case could improve the policies' performance. Given that we trained the policies on only 20 demonstrations, we nevertheless conclude that the policies can generalize to the new testing environments, as the overall performance only drops slightly by 7% reaching 73% on average.

V. CONCLUSION

This work investigated the importance of tactile sensing for performing the dynamic manipulation task of match lighting. The policies were learned from demonstration data obtained through kinesthetic teaching. Our proposed policy learning framework leveraged a flow matching generative model for fast and efficient action generation and a modular multi-modal transformer as policy representation. The experimental results underline that tactile feedback is crucial for learning performant match lighting policies. Across all task variations, the vision+touch policies outperformed vision-only policies, increasing the number of successful

policy rollouts almost by a factor of 3. By analysing the visuotactile policies’ attention weights, we confirmed that tactile observations gain importance during the contact-rich interactions between the match tip and the striker paper. Moreover, we also showed that exploiting the tactile signals during training and employing a masked training procedure can benefit vision-only policies and yield increased success rates. Yet, the improved vision-only policies still cannot reach the performance of the visuotactile policies. Lastly, we showed that the visuotactile policies are robust and can generalize to novel task variations. All findings underline the importance of tactile sensing for obtaining performant policies for successfully solving the dynamic manipulation task of igniting matches. Future work should investigate transferring these findings to other manipulation tasks and further improving the overall performance of the visuotactile policies by, e.g., learning from unsuccessful policy rollouts.

ACKNOWLEDGMENT

We thank Erik Helmut and Rickmer Krohn for helping with the 3D printing and attention analysis. This work has received funding from the German Research Foundation (DFG) Emmy Noether Programme (CH 2676/1-1), the EU’s Horizon Europe project ARISE (Grant no.: 101135959), the AICO grant by the Nexplore/Hochtief Collaboration with TU Darmstadt. This work was partly supported by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany’s Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence “Centre for Tactile Internet with Human-in-the-Loop” (CeTI) of Technische Universität Dresden, and by Bundesministerium für Bildung und Forschung (BMBF) and German Academic Exchange Service (DAAD) in project 57616814 (SECAI, School of Embedded and Composite AI).

REFERENCES

- [1] S. K. Sampath, N. Wang, H. Wu, and C. Yang, “Review on human-like robot manipulation using dexterous hands,” *Cogn. Comput. Syst.*, 2023.
- [2] O. Kroemer, S. Niekum, and G. Konidaris, “A review of robot learning for manipulation: Challenges, representations, and algorithms,” *Journal of machine learning research*, 2021.
- [3] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, “Recent advances in robot learning from demonstration,” *Annual review of control, robotics, and autonomous systems*, 2020.
- [4] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, S. K. S. Ghasemipour, C. Finn, and A. Wahid, “Aloha unleashed: A simple recipe for robot dexterity,” in *Conference on Robot Learning*, 2024.
- [5] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, “Diffusion policy: Visuomotor policy learning via action diffusion,” *IJRR*, 2023.
- [6] M. R. Cutkosky and J. M. Hyde, “Manipulation control with dynamic tactile sensing,” in *International symposium on robotics research*, 1993.
- [7] N. Funk, J. Urain, J. Carvalho, V. Prasad, G. Chalvatzaki, and J. Peters, “Actionflow: Equivariant, accurate, and efficient policies with spatially symmetric flow matching,” *arXiv preprint arXiv:2409.04576*, 2024.
- [8] K. Kronander and A. Billard, “Learning compliant manipulation through kinesthetic and tactile human-robot interaction,” *IEEE transactions on Haptics*, 2013.
- [9] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, “3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing,” in *Conference on Robot Learning*, 2024.

- [10] K. Yu, Y. Han, Q. Wang, V. Saxena, D. Xu, and Y. Zhao, “Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation,” in *Conference on Robot Learning*, 2024.
- [11] R. S. Johansson, “The effects of anesthesia on motor skills,” [Online] - <https://www.youtube.com/watch?v=0LfJ3M3Kn80>, [Accessed 15-12-2024].
- [12] R. T. Chen and Y. Lipman, “Riemannian flow matching on general geometries,” *arXiv preprint arXiv:2302.03660*, 2023.
- [13] Q. Li, O. Kroemer, Z. Su, F. F. Veiga, M. Kaboli, and H. J. Ritter, “A review of tactile information: Perception and action through touch,” *IEEE Transactions on Robotics*, 2020.
- [14] W. Yuan, S. Dong, and E. H. Adelson, “Gelsight: High-resolution robot tactile sensors for estimating geometry and force,” *Sensors*, 2017.
- [15] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, “The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies,” *Soft robotics*, vol. 5, no. 2, pp. 216–227, 2018.
- [16] N. Funk, E. Helmut, G. Chalvatzaki, R. Calandra, and J. Peters, “Eventac: An event-based optical tactile sensor for robotic manipulation,” *IEEE Transactions on Robotics*, 2024.
- [17] S. Dong, D. K. Jha, D. Romeres, S. Kim, D. Nikovski, and A. Rodriguez, “Tactile-rl for insertion: Generalization to objects of unknown geometry,” in *ICRA*, 2021.
- [18] J. Hansen, F. Hogan, D. Rivkin, D. Meger, M. Jenkin, and G. Dudek, “Visuotactile-rl: Learning multimodal manipulation policies with deep reinforcement learning,” in *ICRA*, 2022.
- [19] A. Church, J. Lloyd, N. F. Lepora *et al.*, “Tactile sim-to-real policy transfer via real-to-sim image translation,” in *Conference on Robot Learning*, 2022.
- [20] C. Sferrazza, Y. Seo, H. Liu, Y. Lee, and P. Abbeel, “The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning,” 2023.
- [21] T. Bi, C. Sferrazza, and R. D’Andrea, “Zero-shot sim-to-real transfer of tactile control policies for aggressive swing-up manipulation,” *IEEE Robotics and Automation Letters*, 2021.
- [22] J. Lenz, T. Gruner, D. Palenicek, T. Schneider, I. Pfenning, and J. Peters, “Analysing the interplay of vision and touch for dexterous insertion tasks,” in *CoRL Workshop on Learning Robot Fine and Dexterous Manipulation: Perception and Control*, 2024.
- [23] J. Urain, A. Mandlekar, Y. Du, M. Shafiuallah, D. Xu, K. Fragkiadaki, G. Chalvatzaki, and J. Peters, “Deep generative models in robotics: A survey on learning from multimodal demonstrations,” *arXiv preprint arXiv:2408.04380*, 2024.
- [24] T. Ablett, Y. Zhai, and J. Kelly, “Seeing all the angles: Learning multiview manipulation policies for contact-rich tasks from demonstrations,” in *IROS*, 2021.
- [25] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, “What matters in learning from offline human demonstrations for robot manipulation,” in *Conference on Robot Learning*, 2021.
- [26] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, “Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware,” in *RSS*, 2023.
- [27] D. Mukashv, S. Seitzhan, J. Chumakov, S. Khajikhanov, M. Yergibay, N. Zhaniyar, R. Chibar, A. Mazhitov, M. Rubagotti, and Z. Kappasov, “E-bts: Event-based tactile sensor for haptic teleoperation in augmented reality,” *IEEE Transactions on Robotics*, 2024.
- [28] T. Ablett, O. Limoyo, A. Sigal, A. Jilani, J. Kelly, K. Siddiqi, F. Hogan, and G. Dudek, “Multimodal and force-matched imitation learning with a see-through visuotactile sensor,” *IEEE Transactions on Robotics*, 2024.
- [29] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, “ π_0 : A vision-language-action flow model for general robot control,” *arXiv preprint arXiv:2410.24164*, 2024.
- [30] P. Xu, X. Zhu, and D. A. Clifton, “Multimodal learning with transformers: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [31] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE conference on computer vision and pattern recognition*, 2016.
- [32] A. Vaswani, “Attention is all you need,” *arXiv preprint arXiv:1706.03762*, 2017.
- [33] “Franka Interactive Controllers,” <https://github.com/nbfigueroa/franka-interactive-controllers>, [Accessed 02-09-2024].